


# Lip-Reading Driven Deep Learning Approach for Speech Enhancement

Ahsan Adeel , Mandar Gogate, Amir Hussain, and William M. Whitmer

**Abstract**—This paper proposes a novel lip-reading driven deep learning framework for speech enhancement. The approach leverages the complementary strengths of both deep learning and analytical acoustic modeling (filtering-based approach) as compared to benchmark approaches that rely only on deep learning. The proposed audio-visual (AV) speech enhancement framework operates at two levels. In the first level, a novel deep learning based lip-reading regression model is employed. In the second level, lip-reading approximated clean-audio features are exploited, using an enhanced, visually-derived Wiener filter (EVWF), for estimating the clean audio power spectrum. Specifically, a stacked long-short-term memory (LSTM) based lip-reading regression model is designed for estimating the clean audio features using only temporal visual features (i.e., lip reading), by considering a range of prior visual frames. For clean speech spectrum estimation, a new filterbank-domain EVWF is formulated, which exploits the estimated speech features. The EVWF is compared with conventional spectral subtraction and log-minimum mean-square error methods using both ideal AV mapping and LSTM driven AV mapping approaches. The potential of the proposed AV speech enhancement framework is evaluated under four different dynamic real-world scenarios [cafe, street junction, public transport, and pedestrian area] at different SNR levels (ranging from low to high SNRs) using benchmark grid and ChiME3 corpora. For objective testing, perceptual evaluation of speech quality is used to evaluate the quality of restored speech. For subjective testing, the standard mean-opinion-score method is used with inferential statistics. Comparative simulation results demonstrate significant lip-reading and speech enhancement improvements in terms of both speech quality and speech intelligibility. Ongoing work is aimed at enhancing the accuracy and generalization capability of the deep learning driven lip-reading model, using contextual integration of AV cues, leading to context-aware, autonomous AV speech enhancement.

**Index Terms**—Lip-reading, stacked long-short-term memory, enhanced visually-derived Wiener filtering, context-aware audio-visual speech enhancement, audio-visual ChiME3 corpus.

## I. INTRODUCTION

**S**PEECH enhancement aims to enhance perceived overall speech quality and intelligibility, when noise degrades them significantly. Due to the need for speech enhancement in a wide-range of real-world applications, such as mobile communication, speech recognition, hearing aids etc., several speech enhancement methods have been proposed over the past few decades, ranging from state-of-the-art statistical, analytical, and classical optimization approaches to advanced deep learning based methods. Classic speech enhancement methods are mainly based on audio only processing such as SS [1], audio-only Wiener filtering [2], minimum mean-square error (MMSE) [3], [4], and linear minimum mean square error (LMMSE) etc. Recently, researchers have proposed deep learning based advanced speech recognition [5] and enhancement [6] approaches. However, most of these are based on single channel (audio only) processing, which often perform poorly in adverse conditions, where overwhelming noise is present [7]. Specifically, state-of-the-art audio-only speech processing algorithms make the signal more audible yet remain deficient in restoring intelligibility. Consequently, more advanced brain-inspired, multimodal speech processing algorithms are required that mimic the human contextual AV speech processing mechanism, to selectively amplify the attended talker or filter out acoustic clutter. In this context, next-generation lip-reading driven AV speech processing stands as an emerging technology in computational intelligence to realise more intelligible audio.

Human speech processing is inherently multimodal, where visual cues help to better understand speech. The multimodal nature of speech is well established in literature, where we understand how speech is produced by the vibration of the vocal folds and configuration of the articulatory organs. The correlation between the visible properties of articulatory organs (e.g., lips, teeth, tongue) and speech reception has been previously shown in numerous behavioural studies [8]–[11]. Therefore, the clear visibility of some articulatory organs could be effectively utilized to approximate a clean speech signal out of a noisy audio background. The biggest advantage of using visual cues to extract clean audio features is their inherent noise immunity: visual speech representation always remains unaffected by acoustic noise [12].

In recent literature, extensive research has been carried out to develop multimodal speech processing methods, which has established the importance of multimodal information in speech processing [13]. Researchers have proposed novel audio feature extraction methods [14]–[16], novel visual feature extraction

Manuscript received January 6, 2019; revised April 8, 2019; accepted April 28, 2019. This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Grant No. EP/M026981/1. (CogAVHearing <http://cogavhearing.cs.stir.ac.uk>). (Corresponding author: Ahsan Adeel.)

A. Adeel is with DeepCI, Edinburgh EH16 5XW, U.K., and also with the School of Mathematics and Computer Science, University of Wolverhampton, Wolverhampton WV1 1LY, U.K. (e-mail: ahsan.adeel@deepci.org).

M. Gogate and A. Hussain are with the School of Computing, Edinburgh Napier University, Edinburgh EH11 4DY, U.K. (e-mail: mgo@cs.stir.ac.uk; A.Hussain@napier.ac.uk).

W. M. Whitmer is with the MRC/CSO Institute of Hearing Research Scottish Section, Glasgow G31 2ER, U.K. (e-mail: william.whitmer@nottingham.ac.uk).

Digital Object Identifier 10.1109/TETCI.2019.2917039

methods [17]–[21], fusion approaches (early integration [22], late integration [23], hybrid integration [24]), multi-modal datasets [25], [26], and fusion techniques [24], [27], [28]. Multimodal audiovisual speech processing methods have demonstrated significant performance improvements [29]. Moreover, with the advent of advanced and more sophisticated digital signal processing approaches (in terms of both hardware and software), researchers have shown some ground breaking performance improvements. For example, the authors in [30] proposed a novel deep learning based lip-reading system with 93% accuracy. The authors in [31] proposed a multimodal hybrid a deep neural network (DNN) architecture based on deep convolutional neural network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) network for speech enhancement. Similarly, the authors in [32] proposed an audiovisual speech recognition system, where deep learning algorithms, such as CNN, deep denoising autoencoders, and multistream hidden Markov model (MSHMM) models have been used for visual feature extraction, audio feature extraction, and audiovisual features integration respectively.

Recently, the authors in [33] proposed an audio-visual speech enhancement approach using multimodal deep CNNs. Specifically, they developed an audio-visual deep CNN (AVDCNN) speech enhancement model that integrates audio and visual cues into a unified network model. The proposed AVDCNN approach is structured as an audio-visual encoder-decoder where both audio and visual cues are processed using two separate CNNs, and later fused into a joint network to generate enhanced speech. However, the proposed AVDCNN approach relies only on deep CNN models. For testing, the authors used self-prepared dataset that contained video recordings of 320 utterances of Mandarin sentences spoken by a native speaker (only one speaker). In addition, the used noises include a normal car engine, baby crying, pure music, music with lyrics, siren, one background talker (1T), two background talkers (2T), and three background talkers (3T).

In contrast to [33], our proposed novel approach leverages the complementary strengths of both deep learning and analytical acoustic modelling (filtering based approach). For testing, we used five different speakers including two white males, two white females, and one black male speaker (that fairly satisfies the speaker independence criteria) with real-world dynamic noises in extreme noisy scenarios. Specifically, we propose a novel deep learning based lip-reading regression model and EVWF for speech enhancement. The framework first reads target speakers lips and estimates clean audio features using stacked LSTM model. Next, the estimated clean audio features are fed into the EVWF for speech enhancement. The proposed approach effectively exploits temporal correlation in lip-movements by considering a varying number of prior visual frames. The presented EVWF model is tested with the challenging benchmark AV ChiME3 corpus, in real-world commercially-motivated scenarios, for SNRs ranging from  $-12$  to  $12$  dB. Comparative simulation results demonstrate that the proposed EVWF approach outperforms benchmark audio-only approaches at very low SNRs ( $-12$  dB,  $-6$  dB,  $-3$  dB, and  $0$  dB), and that the improvement is statistically significant at the 95% confidence level.

The four major contributions presented in this paper are:

- 1) Proposed a novel deep learning based lip-reading regression model for speech enhancement applications. Specifically, a stacked LSTM based data-driven model is proposed to approximate the clean audio features using only temporal visual features (i.e. lip reading). In the literature, extensive research has been carried out to model lip reading as a classification problem for speech recognition. In contrast, not much work has been conducted to model lip reading as a regression problem for speech enhancement [30]–[32].
- 2) A critical analysis of the proposed LSTM based lip-reading regression model and its comparison with the conventional multilayer perceptron (MLP) based regression model [34], where LSTM model has shown better capability to learn the correlation between lip movements and speech as compared to conventional MLP models, particularly, when a different number of prior visual frames are considered.
- 3) Addressed limitations of state-of-the-art VWF by presenting a novel EVWF. The proposed EVWF has employed an inverse filter-bank (FB) transformation (i.e. a pseudoinverse of the approximated audio features) for audio power spectrum estimation as compared to the cubic spline interpolation method (that fails to estimate the missing power spectral values when it interpolates the low dimensional vector to high dimensional audio vector and ultimately leads to a poor audio power spectrum estimation). In addition, the proposed EVWF has eliminated the need for voice activity detection (VAD) and noise estimation.
- 4) Evaluated the potential of our proposed speech enhancement framework, exploiting both ideal AV mapping (ideal visual to audio feature mapping with exact clean audio features) and designed stacked LSTM based lip-reading model. The benchmark AV Grid and ChiME3 corpora are employed, with 4 different real-world noise types (cafe, street junction, public transport (BUS), pedestrian area), to objectively and subjectively evaluate the performance of the EVWF.

The rest of the paper is organized as follows: Section II presents the proposed AV speech enhancement framework, designed EVWF, and lip-reading algorithms. Section III presents the employed AV dataset and audiovisual feature extraction methodology. Section IV presents comparative experimental results, including the evaluation of deep learning-driven AV mapping and EVWF based speech enhancement. Finally, Section V concludes this work and proposes some future research directions.

## II. SPEECH ENHANCEMENT FRAMEWORK

The proposed two-level, deep learning based, lip reading driven speech enhancement framework is depicted in Fig. 1. In the first level, a novel lip-reading regression model is effectively utilized to estimate clean audio features using only temporal visual features. In the second level, estimated low dimensional

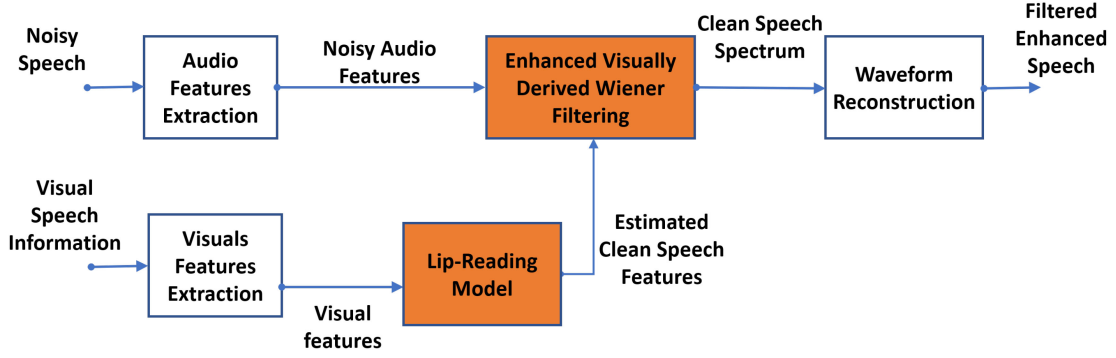


Fig. 1. Proposed lip-reading driven deep learning approach for speech enhancement. The system first estimates clean audio features using a visually derived (i.e. lip reading) speech model. Next, the estimated clean audio features are fed into the proposed enhanced, visually-derived Wiener filter for estimating the clean speech spectrum.

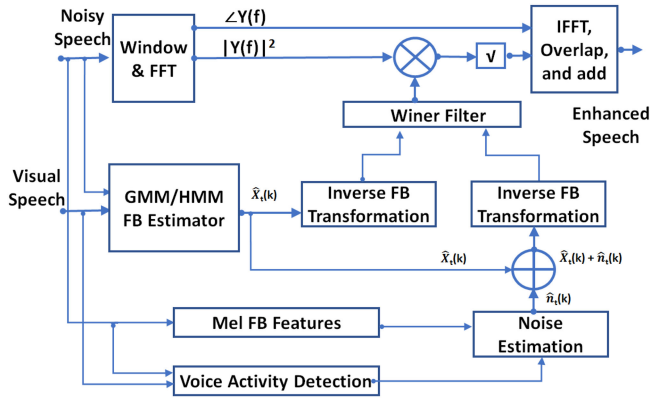


Fig. 2. State-of-the-art visually-derived Wiener filtering [12]. The authors in [12], presented a hidden Markov model-Gaussian mixture model (HMM/GMM) based two-level state-of-the-art VWF for speech enhancement. However, the use of HMM/GMM models for the estimation of clean audio features ( $\hat{X}_t(i)$ ) from visual features and cubic spline interpolation for the approximation of high dimensional clean audio power spectrum from the estimated low dimensional audio features are not optimal choices. The HMM/GMM model suffers from poor generalization and cubic spline interpolation method fails to estimate the missing power spectral values that leads to a poor audio power spectrum estimation.

clean audio features are transformed into a high dimensional clean audio power spectrum using inverse FB transformation to calculate the Wiener filter. Finally, the Wiener filter is applied to the magnitude spectrum of the noisy input audio signal, followed by the inverse fast Fourier transform (IFFT), overlap, and combining processes to produce enhanced magnitude spectrum. The state-of-the-art VWF and designed EVWF are depicted in Fig. 2. and Fig. 3 respectively. It is to be noted that the designed EVWF has addressed the limitations of state-of-the-art VWF [12] by employing an inverse FB transformation (i.e. a pseudoinverse of the approximated audio features) for estimating the audio power spectrum as compared to the cubic spline interpolation method. In addition, it has eliminated the need for VAD and noise estimation.

#### A. Enhanced Visually Derived Wiener Filter

In signal processing, the Wiener filter is a state-of-the-art filter that helps to produce an estimate of a clean audio signal by linear

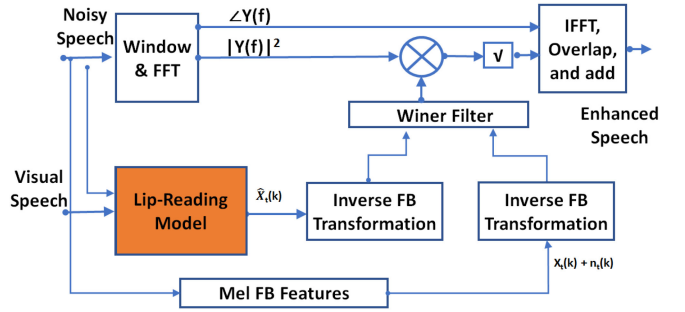


Fig. 3. Proposed enhanced visually-derived Wiener filtering. Note the use of LSTM based FB estimation and an inverse FB transformation for audio power spectrum estimation. The approach addressed both the power spectrum estimation and generalization issues of the state-of-the-art VWF. In addition, it replaced the need for a voice activity detector and noise estimator.

time-invariant (LTI) filtering of an observed noisy audio signal. The frequency domain Wiener Filter is defined as:

$$W(\gamma) = \frac{\psi_a(\gamma)}{\psi_a(\gamma)} \quad (1)$$

where  $\psi_a(\gamma)$  is the noisy audio power spectrum (i.e. clean power spectrum + noisy power spectrum) and  $\psi_a(\gamma)$  is the clean audio power spectrum. The calculation of the noisy audio power spectrum is fairly easy because of the available noisy audio vector. However, the calculation of the clean audio power spectrum is challenging which restricts the use of the Wiener filter widely. Hence, for successful Wiener filtering, it is necessary to acquire the clean audio power spectrum. In this paper, the clean audio power spectrum is calculated using deep learning based lip reading driven speech model.

The FB domain Wiener filter ( $\hat{W}_t^{FB}(k)$ ) is given as [12]:

$$\hat{W}_t^{FB}(k) = \frac{\hat{x}_t(k)}{\hat{x}_t(k) + \hat{n}_t(k)} \quad (2)$$

where  $\hat{x}_t(k)$  is the FB domain lip-reading driven approximated clean audio feature and  $\hat{n}_t(k)$  is the FB domain noise signal. The subscripts  $k$  and  $t$  represents the  $k^{th}$  channel and  $t^{th}$  audio frame.

The lip-reading driven approximated clean audio feature vector  $\hat{x}_t(k)$  is a low dimensional vector. For high dimensional

Wiener filter calculation, it is necessary to transform the estimated low dimensional FB domain audio coefficients into a high dimensional power spectral domain. It is to be noted that the approximated clean audio and noise features ( $\hat{x}_t(k) + \hat{n}_t(k)$ ) are replaced with the noisy audio (a combination of real clean speech ( $x_t(k)$ ) and noise ( $n_t(k)$ )). The low dimension to high dimension transformation can be written as:

$$\hat{W}_{t[N_l, M]}^{FB}(k) = \frac{\hat{x}_t(k)_{[N_l, M]}}{x_t(k)_{[N_l, M]} + n_t(k)_{[N_l, M]}} \quad (3)$$

$$\hat{W}_{t[N_h, M]}^{FB}(k) = \frac{\hat{x}_t(k)_{[N_h, M]}}{x_t(k)_{[N_h, M]} + n_t(k)_{[N_h, M]}} \quad (4)$$

where  $N_l$  and  $N_h$  are the low and high dimensional audio features respectively, and  $M$  is the number of audio frames. The transformation from (3) to (4) is very crucial to the performance of filtering. Therefore, the authors in [12] proposed the use of cubic spline interpolation method to determine the missing spectral values. However, the use of cubic spline interpolation for the approximation of high dimensional clean audio power spectrum from a low dimensional audio FB features is not an optimal approach. The interpolation based method fails to estimate the missing power spectral values. In contrast, this article proposes the use of inverse FB transformation which used the least square pseudo-inverse based method to approximate the optimal high dimensional power spectrum.

The inverse FB domain transformation is calculated as follows:

$$\hat{x}_t(k)_{[N_h, M]} = \hat{x}_t(k)_{[N_l, M]} * \alpha_x \quad (5)$$

$$n_t(k)_{[N_h, M]} = n_t(k)_{[N_l, M]} * \alpha_n \quad (6)$$

$$\alpha_x = \alpha_n = (\phi_m(k)^T \phi_m(k))^{-1} \phi_m(k)^T \quad (7)$$

$$\phi_m(k) = \begin{cases} 0 & k < f_{b_{mf-1}} \\ \frac{k - f_{b_{mf-1}}}{f_{b_{mf}} - f_{b_{mf-1}}} & f_{b_{mf-1}} \leq k \leq f_{b_{mf}} \\ \frac{f_{b_{mf+1}} - k}{f_{b_{mf+1}} - f_{b_{mf}}} & f_{b_{mf}} \leq k \leq f_{b_{mf+1}} \\ 0 & k > f_{b_{mf+1}} \end{cases}$$

where  $f_{b_{mf}}$  are the boundary points of the filters and corresponds to the  $k^{th}$  coefficient of the  $k$ -points DFT. The boundary points  $f_{b_{mf}}$  are calculated using:

$$f_{b_{mf}} = \left( \frac{K}{F_{smp}} \cdot f_{cmf} \right) \quad (8)$$

where  $f_{cmf}$  is the mel scale frequency

After substitutions, the obtained  $k$ -bin Wiener filter (4) is then applied to the magnitude spectrum of the noisy audio signal (i.e.  $|Y_t(k)|$ ) to estimate the enhanced magnitude audio spectrum ( $|\hat{X}_t(k)_{[N_h, M]}|$ ). The enhanced magnitude audio spectrum is given as:

$$|\hat{X}_t(k)| = |Y_t(k)| |\hat{W}_{t[N_h, M]}| \quad (9)$$

The acquired time-domain enhanced speech signal (9) is then followed by the IFFT, overlap, and combining processes.

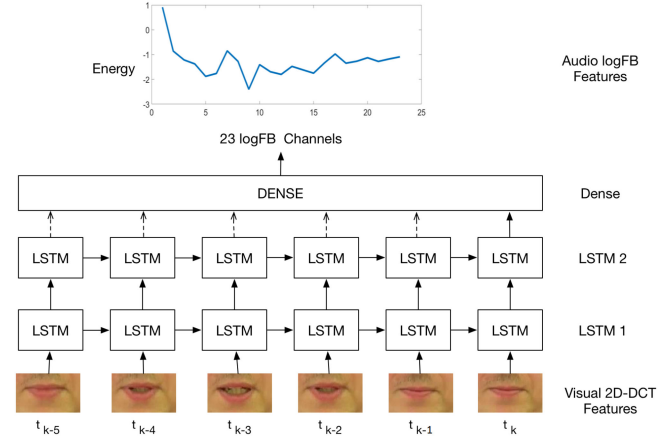


Fig. 4. Stacked long-short-term memory based lip reading regression model: Example of five prior visual frames (taking into account the current visual frame  $t_k$  as well as the temporal information of previous visual frames  $t_{k-1}$ ,  $t_{k-2}$ ,  $t_{k-3}$ ,  $t_{k-4}$ ,  $t_{k-5}$ ).

### B. Lip Reading Regression Model

This section describes the LSTM based lip reading regression model summarised in Fig. 4. LSTM was originally proposed in [35] by Sepp Hochreiter and Jrgen Schmidhuber. The LSTM network consists of input layer, two LSTM layers, and output dense layer. Visual features of time instance  $t_k, t_{k-1}, \dots, t_{k-5}$  ( $k$  is the current time instance and 5 is the number of prior visual frames) were feeded into the stacked LSTM layers. The lower LSTM layer has 250 cells, which encoded the input and passed its hidden state to the second LSTM layer, which has 300 cells. The output of the second LSTM layer was then feeded into the fully connected (dense) layer which has total 23 neurons with linear activation function. In this architecture, the input at layer  $k$  is the value of hidden state  $h_t$  computed by layer  $k-1$ . The stacked LSTM architecture was trained with the objective to minimise the mean squared error (MSE) between the predicted and the actual audio features. The MSE (10) between the estimated audio logFB features and clean audio features was minimised using stochastic gradient decent algorithm and RMSProp optimiser. RMSprop is an adaptive learning rate optimizer which divides the learning rate by moving average of the magnitudes of recent gradients to make learning more efficient. Moreover, to reduce the overfitting, dropout (of 0.25) was applied after every LSTM layer. The MSE cost function  $C(a_{estimated}, a_{clean})$  can be written as:

$$C(a_{estimated}, a_{clean}) = \sum_{i=1}^n 0.5(a_{estimated}(i) - a_{clean}(i))^2 \quad (10)$$

where  $a_{estimated}$  and  $a_{clean}$  are the estimated and clean audio features respectively.

### C. Input & Output Preprocessing

Both LSTM and MLP networks ingest visual discrete cosine transform (DCT) features of time instance  $t_n, t_{n-1}, \dots, t_{n-k}$ , where  $n$  is the current time instance and  $k$  is the number of prior visual frames as shown in Fig. 4. The Input layer of the



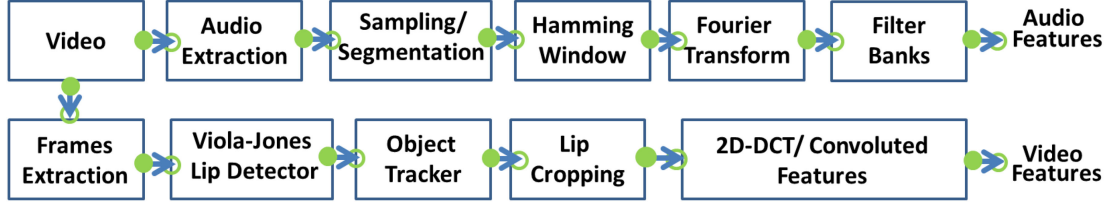


Fig. 5. Audiovisual dataset generation procedure.

TABLE I  
SUMMARY OF SENTENCES FROM THE GRID CORPUS

Speaker ID	Grid ID	Total	Full		Aligned	
			Removed	Used	Removed	Used
Speaker 1	S1	1000	11	989	11	989
Speaker 2	S15	1000	164	836	164	836
Speaker 3	S26	1000	16	984	71	929
Speaker 4	S6	1000	9	991	9	991
Speaker 5	S7	1000	11	989	11	989

networks are organised such that at  $k^{th}$  time step LSTM/MLP receives temporal input. The output of the dense layer is logFB audio feature.

### III. DATASET AND AUDIOVISUAL FEATURE EXTRACTION

#### A. Dataset

In this paper, a well-established Grid [25] and ChiME3 [36] corpora are used. The clean Grid videos are mixed with ChiME3 noises (cafe, street junction, public transport (BUS), pedestrian area) for SNRs ranging from  $-12$  to  $12$  dB to develop a new AV ChiME3 corpus. The preprocessing includes sentence alignment and incorporation of prior visual frames. Sentence alignment is performed to remove the silence time from the video and prevent the model from learning redundant or insignificant information. Preprocessing enforced the model to learn the correlation between the spoken word and corresponding visual representation, rather than over learning the silence. Secondly, prior multiple visual frames are used to incorporate temporal information to improve mapping between visual and audio features. The Grid corpus comprised of 34 speakers, each speaker reciting 1000 sentences, and each sentence consists of a six-word sequence of the form indicated in [25] and [36] e.g., *bin blue at A 9 again, lay green by minus W zero now* etc. These sentences comply with International Organization for Standardization (ISO) intelligibility testing recommendations and fall in the category of nonsense words with a difficulty level of 5 (hardest) [37]. Out of 34 speakers, a subset of 5 speakers is selected (two white females, two white males, and one black male) with total 900 command sentences each. The subset fairly ensures the speaker independence criteria. A summary of the acquired visual dataset is presented in Table I, where the full and aligned sentences, total number of sentences, used sentences, and removed sentences are clearly defined. The audio and visual features extraction procedure is depicted in Fig. 5.

#### B. Audio Feature Extraction

The audio features are extracted using widely used log-FB vectors [34]. For log-FB vectors calculation, the input audio

signal is sampled at 50kHz and segmented into  $N$  16ms frames with 800 samples per frame and 62.5% increment rate. Afterwards, a hamming window and Fourier transformation is applied to produce 2048-bin power spectrum. Finally, a 23-dimensional log-FB is applied, followed by the logarithmic compression to produce 23-D log-FB signal.

#### C. Visual Feature Extraction

The visual features are extracted from the Grid Corpus videos recorded at 25 fps using a 2D-DCT based standard and widely used visual feature extraction method. Firstly, the video files are processed to extract a sequence of individual frames. Secondly, a Viola-Jones lip detector [38] is used to identify a lip-region by defining the Region-of-Interest (ROI) in terms of bounding box. Object detection is performed using Haar feature-based cascade classifiers. The method is based on machine learning where cascade function is trained with positive and negative images. Finally, the object tracker [39] is used to track lip regions across the sequence of frames. The visual extraction procedure produced a set of corner points for each frame, where lip regions are then extracted by cropping the raw image. In addition, to ensure good lip tracking, each sentence is manually validated by inspecting a few frames from each sentence. The aim of manual validation is to delete those sentences in which lip regions are not correctly identified [34]. Lip tracking optimization lies outside the scope of the present work. The development of a full autonomous system and its testing on challenging datasets is in progress.

In the final stage of visual features extraction, the 2D-DCT of a lip region is calculated to produce vectors of pixel intensities. In order, to produce the final frame vector, first 50 components are vectorized in a zigzag order and then the DCT features are interpolated to match the equivalent audio sequence. It is to be noted that videos are recorded at 25 fps and corresponding audio files are produced at 75 vectors per second (VPS). Therefore, the produced visual vectors are upsampled to match the 75 VPS rate. This is performed by copying each visual vector three times and mapping to three consecutive audio frames (e.g.  $V1, V1, V1 \rightarrow A1, A2, A3$ ).

## IV. EXPERIMENTAL RESULTS

#### A. Methodology

For audiovisual mapping and estimation of clean audio features, both MLP and LSTM models are used. Firstly, in Section IV-B, an MLP based learning model is trained and validated for initial data analysis [31]. The dataset analysis with

TABLE II  
SUMMARY OF TRAIN, TEST, AND VALIDATION SENTENCES FROM GRID CORPUS

Speakers	Train	Validation	Test	Total
1	692	99	198	989
2	585	84	167	836
3	650	93	186	929
4	693	99	199	991
5	692	99	198	989
All	3312	474	948	4734

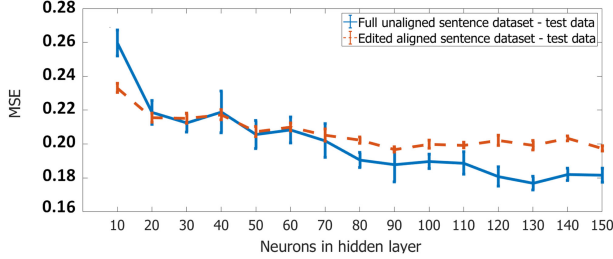


Fig. 6. MSE median test data results of using full (solid blue line) and aligned (red dashed line) sentences for different MLP hidden layer sizes. MLP is used for initial data analysis to evaluate unaligned and aligned sentences. The analysis revealed that the full sentence learning model overlearns silence due to over representation.

MLP revealed that the selected dataset possesses enough speakers variability to demonstrate the potential of our proposed approach and that it is optimal to use aligned sentences. In Section IV-C, both MLP and LSTM models are tested and compared using multiple visual frames. A subset of the dataset is used to train the neural network (80% training dataset) and rest of the data (20%) is used to test and validate the performance of the trained neural network in the face of new contexts (10% validation, 10% testing). Table II summarizes the Train, Test, and Validation sentences. For generalization testing, the proposed framework is trained on SNRs ranging from  $-10$  to  $12$  dB, and tested on  $-12$  dB SNR. Neural network performances are evaluated using MSE with the goal to achieve the least possible MSE. The estimated lip-reading driven audio features are exploited by the designed novel EVWF for speech enhancement. The speech enhancement of our proposed EVWF is compared with the state-of-the-art SS and Log MMSE based audio-only speech enhancement approaches. For noisy speech, clean utterances were mixed with noisy backgrounds (randomly chosen from ChiME3 noises) to produce noisy utterances of SNRs ranging from  $-12$  dB to  $12$  dB.

### B. Initial Data Analysis With MLP

1) *Sentence Length and 1 to 1 Visual to Audio Mapping Evaluation Using MLP*: In this subsection, the use of aligned sentences is justified. For sentence length evaluation, full and aligned sentences of all five speakers with 900 sentences each are used. Audio and visual features are extracted and shuffled randomly. Sentence length evaluation results are shown in Fig. 6. The MLP model is tested with a different number of neurons and layers (e.g. 1 layer with 60 neurons, 2 layers with 60 neurons, 2 layers with 120 neurons, 3 layers with 150 neurons etc.) It can be seen that the learning model performs better with full sentences as compared to the aligned sentences. However, closer

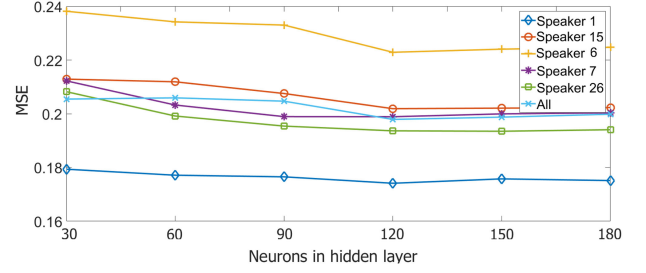


Fig. 7. MSE test data results of using individual and combined speakers (for 1 to 1 audiovisual mapping). MLP is used for initial data analysis to evaluate the fairness of 5 speakers subset and the speaker independence criteria. The simulation results revealed that different speakers have achieved different MSEs due to different speakers articulation (that leads to good or bad audiovisual mapping).

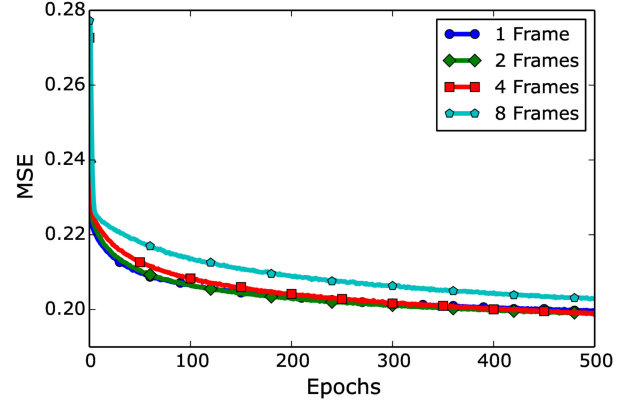


Fig. 8. AV mapping: MLP validation results for different visual frames - All Speakers. The figure presents an overall behaviour of an MLP model when contextual information (i.e. previous frames) is added. It is to be noted that MLP failed to acquire significant performance improvement upon contextual information integration.

inspection revealed that the learning model with full sentences over learns silence due to silence representation. In addition, the full sentence learning model works as a VAD, where it effectively distinguishes between the silent and speech frames. However, in this article, we aim to learn the relationship between speech and visuals for speech mapping instead of modelling VAD. Therefore, the aligned dataset is of more relevance even if it shows a higher MSE.

2) *Evaluation of Speaker Independence Criteria*: In this subsection, the aim is to justify the use of a 5 speaker subset for evaluating the proposed approach. Fig. 7 shows both speaker dependent and independent audiovisual mapping results with MLP. MLP with different number of neurons and layers is tested (e.g. 1 layer with 60 neurons, 2 layers with 60 neurons, 2 layers with 120 neurons, 3 layers with 150 neurons etc.) It can be seen that different speakers have achieved different MSE results. The variations in MSE is because of different speakers articulation that leads to good or bad audiovisual mapping; hence, satisfying speaker independence criteria.

### C. Multiple Visual Frames to Audio Mapping Using MLP and LSTM

In multiple visual frames to audio mapping, multiple prior frames are used (ranging from 1 visual frame to 27 prior

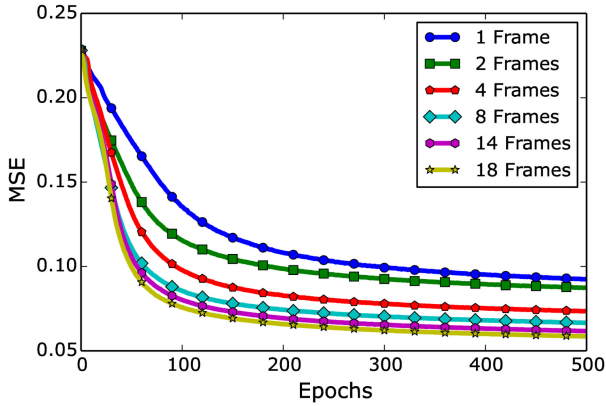


Fig. 9. AV mapping: Stacked LSTM validation results for different visual frames - All Speakers. The figure presents an overall behaviour of an LSTM model when contextual information (i.e. previous frames) is added. It is to be noted that LSTM better exploited the temporal correlation as compared to MLP. However, LSTM saturates at 18 prior visual frames.

TABLE III

MLP VS. LSTM: TRAINING AND TESTING ACCURACY COMPARISON FOR DIFFERENT VISUAL FRAMES

Visual Frames	LSTM		MLP	
	$MSE_{train}$	$MSE_{test}$	$MSE_{train}$	$MSE_{test}$
1	0.092	0.104	0.199	0.204
2	0.087	0.097	0.199	0.202
4	0.073	0.085	0.198	0.200
8	0.066	0.082	0.202	0.204
14	0.061	0.080	0.210	0.218
18	0.058	0.078	0.217	0.209

visual frames). Simulation results are shown in Figs. 8 and 9 and Table III. Training is performed with six different aligned datasets (i.e 1, 2, 4, 8, 14, and 18 prior visual frames). The datasets varied in total number of visual frames (including current and prior visual frames). The multiple prior visual frames correspondingly increased the total number of inputs. However, the output dimensions remained same. It can be seen that by moving from 1 visual frame to 18 visual frames, a significant performance improvement could be achieved. The LSTM model with 1 visual frame achieved the MSE of 0.092, whereas with 18 visual frames, the model achieved the least MSE of 0.058. In contrast, the MLP based lip reading model could only achieve the MSE of 0.199 and 0.209 with 1 and 18 visual frames respectively. It is to be noted that MLP remained deficient in achieving low MSE, because MLP architecture lacks the capability of exploiting prior visual frames for better learning. In contrast, LSTM based learning model exploited the temporal information (i.e. prior visual frames) effectively and showed a consistent reduction in MSE from 1 to 18 visual frames. This is mainly because of its inherent recurrent architectural property and its ability to retain state over long time spans by using cell gates. The estimated log-FB vectors based on visual inputs are shown in Fig. 10, with both MLP and LSTM models. LSTM's enhanced visual to audio mapping as compared to MLP is evident.

#### D. Speech Enhancement

1) *Objective Test:* For objective testing, perceptual evaluation of speech quality (PESQ) is used to evaluate the quality of restored speech using AV ChiME3 corpus. The PESQ

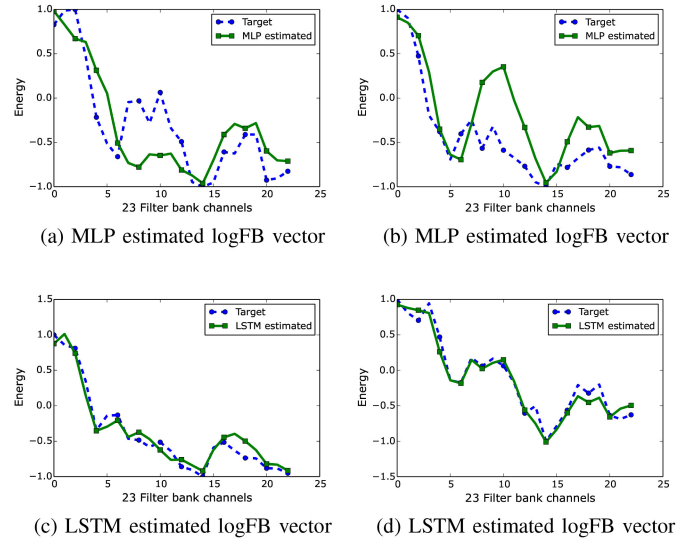


Fig. 10. Estimated clean audio features using 14 prior visual frames.

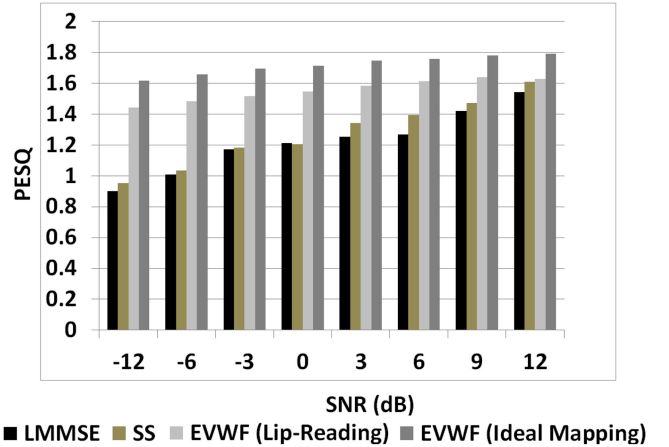


Fig. 11. PESQ with ChiME3 Noises (cafe, street junction, public transport (BUS), pedestrian area). It can be seen that, at low SNR levels, EVWF significantly outperformed both SS and LMMSE based speech enhancement methods. The low PESQ score with ChiME3 corpus, particularly at high SNRs, can be attributed to the nature of the ChiME3 noise, characterized by spectro-temporal variation, potentially reducing the ability of enhancement algorithms to restore the signal.

score is computed as a linear combination of the average disturbance value and the average asymmetrical disturbance values. The PESQ score ranges from  $-0.5$  to  $4.5$  corresponding to low to high speech quality. The PESQ scores for EVWF, SS, and LMMSE for different SNRs are depicted in Fig. 11. It can be seen that, at low SNR, EVWF significantly outperformed both SS [40] and LMMSE [4] based speech enhancement methods. The low PESQ score with ChiME3 corpus, particularly at high SNRs, can be attributed to the nature of the ChiME3 noise. The latter is characterized by spectro-temporal (ST) variation, potentially reducing the ability of enhancement algorithms to restore the signal. Fig. 12 displays the spectrogram of a randomly selected utterance from AV ChiME3 corpus, where the performance of EVWF at very low SNR ( $-12$ db) is evident. It is to be noted that for generalization testing,  $-12$  dB SNR utterances were not included in the training dataset.



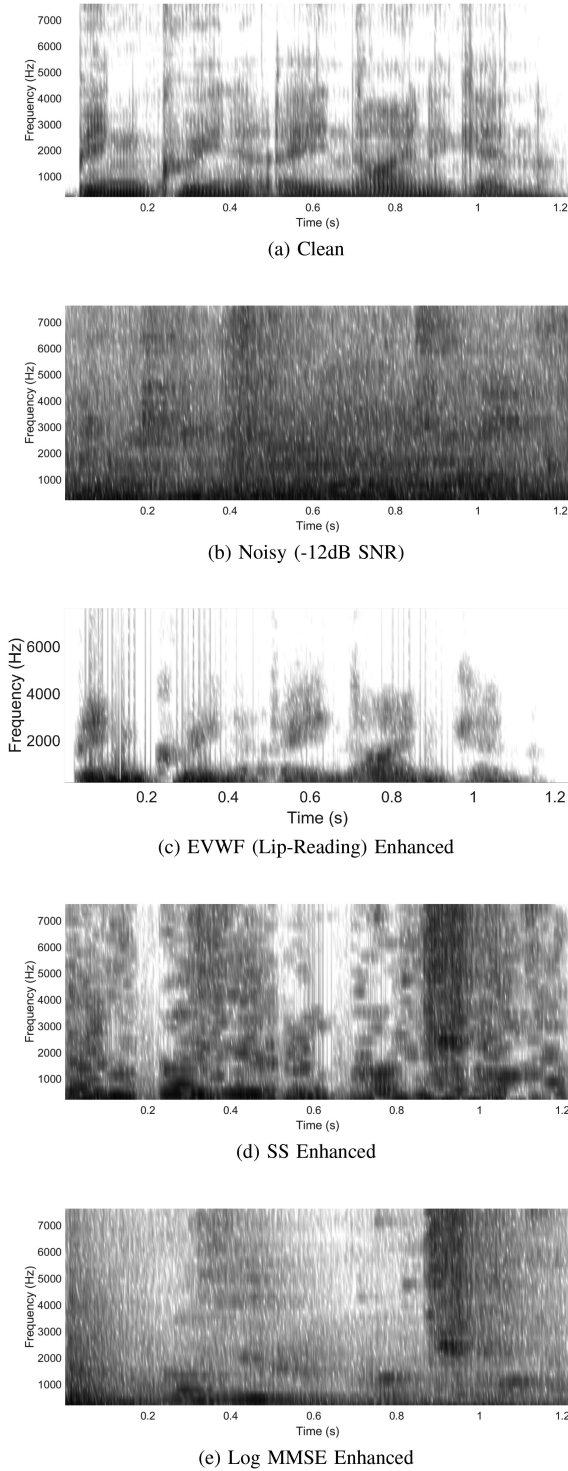


Fig. 12. Spectrogram of a random utterance from AV ChiME3 corpus: (a) Clean (b) Noisy ( $-12$  dB SNR) (c) EVWF (Lip-Reading) Enhanced (d) Spectral Subtraction Enhanced (e) Log MMSE Enhanced.

Whilst state-of-the-art audio-only speech enhancement algorithms such as SS can be effective for filtering noisy speech in stationary conditions, the cleaner processed speech suffers from low intelligibility. Beamforming with multiple microphones can significantly improve intelligibility however this technique is

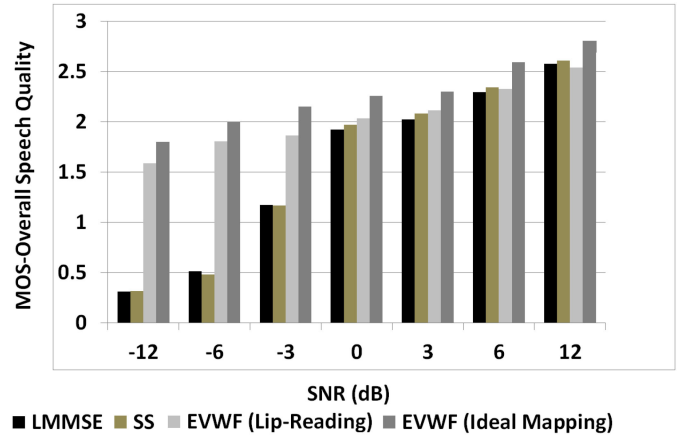


Fig. 13. MOS for overall speech quality with ChiME3 Noises (cafe, street junction, public transport (BUS), pedestrian area). It can be seen that, EVWF significantly outperformed both SS and LMMSE based speech enhancement methods at low SNR.

also difficult to implement in an unpredictable noise environment [41]. In contrast, AV speech enhancement offers consistent intelligibility gains due to its ability to deliver phonetic information obliterated in the masked region. However, the intelligibility is also governed by information masking (IM), that is, the degree by which the auditory system segregates the ST regions which are dominated by the speech from those that are background dominated. In noisy backgrounds, IM is amplified by even mild hearing impairment that leads to a large speech intelligibility loss [41]. Schwartz *et al.* [42] presented a study showing how visual cues can supplement auditory grouping cues, giving a signal that directs attention to the ST regions dominated by the target source [41]. It is believed that AV speech enhancement algorithms are able to mimic the IM releasing function of visual cues. For example, the algorithm would be able to use the visual information to direct audio signal processing to amplify the speech signal components and suppress the noise components [41].

2) *Subjective Listening Tests:* To examine the effectiveness of the proposed EVWF, subjective listening tests were conducted in terms of MOS with self-reported normal-hearing listeners using AV ChiME3 corpus. The listeners were presented with a single stimulus (i.e. enhanced speech only) and were asked to rate the re-constructed speech on a scale of 1 to 5. The five rating choices were: (5) Excellent (when the listener feels unnoticeable difference compared to the target clean speech) (4) Good (perceptible but not annoying) (3) Fair (slightly annoying) (2) Poor (annoying), and (1) Bad (very annoying). The EVWF is compared with two state-of-the-art speech enhancement methods (SS and LMMSE). A total of 10 listeners took part in the evaluation session. The clean speech signal was corrupted with ChiME3 noises (at SNRs of  $-12$  dB,  $-6$  dB,  $-3$  dB,  $0$  dB,  $3$  dB,  $6$  dB, and  $12$  dB). Fig. 13 depicts the performances of three different speech enhancement methods in terms of MOS with ChiME3 noises. In addition, the performance of proposed stack LSTM based lip-reading model is compared with the ideal AV mapping, showing good quality estimation. It can be seen that



TABLE IV

THE RESULTS OF THE T-TEST AT 5% SIGNIFICANCE LEVEL (MOS)-CHiME3: COMPARISON OF EVWF WITH SS. THE PROPOSED EVWF UNDER REAL NOISY ENVIRONMENTS SUGGESTS THAT THE PROPOSED AV APPROACH OUTPERFORMS BENCHMARK SS APPROACH AT LOW SNRS, AND THE IMPROVEMENT IS STATISTICALLY SIGNIFICANT AT THE 95% CONFIDENCE LEVEL

SNRs	p-value	$H_0$ : Null hypothesis
-12dB	5.23E-05	(+)
-6dB	6.23E-05	(+)
-3dB	6.61E-05	(+)
0dB	0.2218	(-)
3dB	0.3795	(-)
6dB	0.9333	(-)
12dB	0.6102	(-)

TABLE V

THE RESULTS OF THE T-TEST AT 5% SIGNIFICANCE LEVEL (MOS)-CHiME3: COMPARISON OF EVWF WITH LMMSE. THE PROPOSED EVWF UNDER REAL NOISY ENVIRONMENTS SUGGESTS THAT THE PROPOSED AV APPROACH OUTPERFORMS BENCHMARK LMMSE APPROACH AT LOW SNRS, AND THE IMPROVEMENT IS STATISTICALLY SIGNIFICANT AT THE 95% CONFIDENCE LEVEL

SNRs	p-value	$H_0$ : Null hypothesis
-12dB	4.16E-05	(+)
-6dB	1.26E-04	(+)
-3dB	4.93E-04	(+)
0dB	0.1149	(-)
3dB	0.2038	(-)
6dB	0.7556	(-)
12dB	0.7183	(-)

the proposed AV approach significantly outperformed benchmark Audio-only approaches at low SNRs. At high SNRs, for the case of (spectro-temporally) correlated ChiME3 noises, the AV performed comparably to Audio-only.

To further determine the significance of our results, we compared the performance of our proposed EVWF with state-of-the-art speech enhancement algorithms, using t-tests at a significance level of 0.05, following statistical analysis with the t-test approach presented in [43]. For each of the pair-wise comparisons, the null hypothesis  $H_0: \mu_1 = \mu_2$  is defined and tested (whether MOS for two methods is significantly different or not for each SNR). The results of the t-test at the level of 0.05 significance are presented in Tables IV (SS vs. EVWF) and V (Log-MMSE vs. EVWF) for AV ChiME3 dataset. It can be seen that the proposed EVWF significantly outperformed both SS and LMMSE at low SNRs ( $-12$  dB,  $-6$  dB, and  $-3$  dB).

## V. DISCUSSION AND CONCLUSION

In this paper, a novel LSTM based lip-reading regression model is first developed, as part of our proposed audio-visual speech enhancement framework. In addition, a novel filter-bank-domain EVWF is formulated and integrated with a lip reading model, and compared with SS and LMMSE methods. The proposed EVWF employs an inverse filter-bank transformation for audio power spectrum estimation, as compared to the cubic spline interpolation method used by the state-of-the-

art VWF model. In addition, it eliminates the need for VAD and noise estimation. Performance evaluation of the lip-reading model demonstrates LSTM's enhanced capability to estimate clean audio features as compared to feed forward neural network models, especially when a different number of prior visual frames are considered. Comparative performance evaluation under real noisy environments suggests that the proposed AV approach outperforms benchmark audio-only approaches at low SNRs, and that the improvement is statistically significant at the 95% confidence level. At high SNRs, performance is still comparable to conventional speech enhancement approaches. The aforementioned limitation of our approach (i.e visual cues become fairly less effective at high SNRs for speech enhancement) leads us to propose future development of a more optimal, context-aware AV system, that can effectively account for different noisy conditions and contextually utilize both visual and noisy audio features. Our ongoing and future work thus aims to develop a context-aware AV integration algorithm to better deal with different noisy environments, and further enhance the accuracy and generalization capability of the current deep learning driven lip-reading model. For generalization testing, this study utilized an adequate (initial) subset of 5 diverse subjects (two white females, two white males, and one black male) with a total of 900 command sentences per speaker. The subset satisfies the speaker independence criteria depicted in Fig. 7, where different speakers can be seen to have achieved different MSEs due to different speakers articulation (that in turn, leads to good or bad AV mapping). The discussion about what constitutes an adequate number of speakers, including testing different datasets, noises etc., requires more work, and will be addressed as a separate topic in a future paper. If we compare our subjective results with recent benchmark works, e.g. Hou *et al.*, 2018 [33], we have conducted more extensive testing in terms of number of speakers and benchmark speech datasets (e.g. Grid and ChiME) at extremely challenging low SNR levels (up to  $-12$  dB). The focus of this work was to demonstrate the potential of our proposed novel AV approach. Our contribution is a first of its kind, in that it leverages the complementary strengths of deep learning and analytical acoustic modelling (filtering based) approaches. In future, we intend to investigate the performance of our EVWF under more realistic scenarios, including generalization testing with unfamiliar speakers, use of novel visual features, and deep speech enhancement components in [41]. Whilst the preliminary comparative results reported in this paper should be taken with care, they demonstrate the potential and capabilities of our developed deep learning-driven AV approach. There is a need for further extensive, comparative evaluation against other benchmark speech enhancement approaches, using a range of real, noisy AV corpora. We are currently recording the latter in real conversational settings, and plan to make these available, as new benchmark resources, to the multidisciplinary speech research community. Ongoing and future work also involves developing a novel open framework for evaluating AV algorithms, including new objective measures and speech intelligibility models, such as new AV speech-in-noise (AV-SIN) listening tests to evaluate real-time AV hearing aids.

## ACKNOWLEDGMENT

In accordance with EPSRC policy, all experimental data used in the project simulations are available at <http://hdl.handle.net/11667/81>. The authors would like to gratefully acknowledge Dr. A. Abel from Xi'an Jiaotong-Liverpool University for building and providing the audio-visual dataset and the authors also would like to acknowledge R. Marxer and J. Barker from the University of Sheffield, R. Watt from the University of Stirling, and P. Derleth from Sonova, AG, Staefa, Switzerland for their contributions. The authors would also like to thank K. Dashtipour for conducting MOS test. Finally, we gratefully acknowledge the support of NVIDIA Corporation for donating the Titan X Pascal GPU for this research.

## REFERENCES

- [1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Aug. 1979.
- [2] W. K. Pratt, "Generalized wiener filtering computation techniques," *IEEE Trans. Comput.*, vol. 100, no. 7, pp. 636–641, Jul. 1972.
- [3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
- [4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.
- [5] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
- [6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.
- [7] J. Benesty, J. Chen, and E. A. Habets, *Speech Enhancement in the STFT Domain*. Berlin, Germany: Springer, 2011.
- [8] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoustical Soc. America*, vol. 26, no. 2, pp. 212–215, 1954.
- [9] Q. Summerfield, "Use of visual information for phonetic perception," *Phonetica*, vol. 36, no. 4/5, pp. 314–331, 1979.
- [10] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.
- [11] M. L. Patterson and J. F. Werker, "Two-month-old infants match phonetic information in lips and voice," *Developmental Sci.*, vol. 6, no. 2, pp. 191–196, 2003.
- [12] I. Almajai and B. Milner, "Visually derived wiener filters for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1642–1651, Aug. 2011.
- [13] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proc. IEEE*, vol. 103, no. 9, pp. 1635–1653, Sep. 2015.
- [14] A. Ezeiza, K. L. de Ipiña, C. Hernández, and N. Barroso, "Enhancing the feature extraction process for automatic speech recognition with fractal dimensions," *Cogn. Comput.*, vol. 5, no. 4, pp. 545–550, 2013.
- [15] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Low-variance multitaper mel-frequency cepstral coefficient features for speech and speaker recognition systems," *Cogn. Comput.*, vol. 5, no. 4, pp. 533–544, 2013.
- [16] S. Nisar, M. Tariq, A. Adeel, M. Gogate, and A. Hussain, "Cognitively inspired feature extraction and speech recognition for automated hearing loss testing," *Cogn. Comput.*, pp. 1–14, 2019.
- [17] C.-h. Chen, *Pattern Recognition and Artificial Intelligence*. Amsterdam, The Netherlands: Elsevier, 2013.
- [18] A. Waibel and K.-F. Lee, *Readings in Speech Recognition*. San Mateo, CA, USA: Morgan Kaufmann, 1990.
- [19] A. Bundy and L. Wallen, "Linear predictive coding," in *Catalogue of Artificial Intelligence Tools*. Berlin, Germany: Springer, 1984, pp. 61.
- [20] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image Vis. Comput.*, vol. 32, no. 9, pp. 590–605, 2014.
- [21] C. Sui, R. Togneri, S. Haque, and M. Bennamoun, "Discrimination comparison between audio and visual features," in *Proc. Conf. Rec. 46th Asilomar Conf. Signals, Syst. Comput.*, 2012, pp. 1609–1612.
- [22] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 11, 2002, Art. no. 783042.
- [23] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 399–402.
- [24] Z. Wu, L. Cai, and H. Meng, "Multi-level fusion of audio and visual features for speaker identification," in *Proc. Int. Conf. Biometrics*, 2006, pp. 493–499.
- [25] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoustical Soc. America*, vol. 120, no. 5, pp. 2421–2424, 2006.
- [26] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 2, pp. II–2017–II 2020.
- [27] G. Iyengar and H. J. Nock, "Discriminative model fusion for semantic concept detection and annotation in video," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 255–258.
- [28] A. K. Noulas and B. J. Kröse, "EM detection of common origin of multimodal cues," in *Proc. 8th Int. Conf. Multimodal Interfaces*, 2006, pp. 201–208.
- [29] F. Berthommier, "A phonetically neutral model of the low-level audio-visual interaction," *Speech Commun.*, vol. 44, no. 1, pp. 31–41, 2004.
- [30] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," *IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2017.
- [31] Z. Wu, S. Sivadas, Y. K. Tan, M. Bin, and R. S. M. Goh, "Multi-modal hybrid deep neural network for speech enhancement," 2016, arXiv:1606.04750.
- [32] K. Noda, Y. Yamaguchi, K. Nakada, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Appl. Intell.*, vol. 42, no. 4, pp. 722–737, 2015.
- [33] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.
- [34] A. Abel *et al.*, "A data driven approach to audiovisual speech mapping," in *Proc. Int. Conf. Brain Inspired Cogn. Syst.*, Springer, 2016, pp. 331–342.
- [35] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proc. 9th Int. Conf. Artif. Neural Netw.*, 1999, pp. 850–855.
- [36] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 504–511.
- [37] M. Vainio, A. Suni, H. Järveläinen, J. Järvelä, and V.-V. Mattila, "Developing a speech intelligibility test based on measuring speech reception thresholds in noise for english and finnish," *J. Acoustical Soc. America*, vol. 118, no. 3, pp. 1742–1750, 2005.
- [38] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 1, pp. 1.511–518.
- [39] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis.*, vol. 77, no. 1/3, pp. 125–141, 2008.
- [40] S. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1979, vol. 4, pp. 200–203.
- [41] A. Hussain *et al.*, "Towards multi-modal hearing aid design and evaluation in realistic audio-visual settings: Challenges and opportunities," in *Proc. 1st Int. Workshop Challenges Hearing Assistive Technol.*, 2017.
- [42] J.-L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: Evidence for early audio-visual interactions in speech identification," *Cognition*, vol. 93, no. 2, pp. B69–B78, 2004.
- [43] S. K. Wajid, A. Hussain, and K. Huang, "Three-dimensional local energy-based shape histogram (3D-lesh)-based feature extraction—a novel technique," *Expert Syst. Appl.*, vol. 112, pp. 388–400, 2017.
- [44] A. Adeel, M. Gogate, and A. Hussain, "Contextual deep learning based audio-visual switching for speech enhancement in real-world environments," *Informat. Fusion*, to be published, 2019.